Optimistic Bounds for Multi-output Prediction

Henry W.J. Reeve [1] **Ata Kabán** [1]

Abstract

We investigate the challenge of multi-output learning, where the goal is to learn a vector-valued function based on a supervised data set. This includes a range of important problems in Machine Learning including multi-target regression, multi-class classification and multi-label classification. We begin our analysis by introducing the self-bounding Lipschitz condition for multi-output loss functions, which interpolates continuously between a classical Lipschitz condition and a multi-dimensional analogue of a smoothness condition. We then show that the self-bounding Lipschitz condition gives rise to optimistic bounds for multi-output learning, which attain the minimax optimal rate up to logarithmic factors. The proof exploits local Rademacher complexity combined with a powerful minoration inequality due to Srebro, Sridharan and Tewari. As an application we derive a state-of-the-art generalisation bound for multi-class gradient boosting.

1. Introduction

Multi-output prediction represents an important class of problems that includes multi-class classification (Crammer & Singer, 2001), multi-label learning (Tsoumakas & Katakis, 2007; Zhang & Zhou, 2013), multi-target regression (Borchani et al., 2015), label distribution learning (Geng, 2016), structured regression (Cortes et al., 2016) and others, with a wide range of practical applications (Xu et al., 2019).

Our objective is to provide a general framework for establishing guarantees for multi-output prediction problems. A fundamental challenge in the statistical learning theory of multi-output prediction is to obtain bounds that allow for (i) favourable convergence rate with the sample size, and (ii) favourable dependence of the risk on the dimensionality of the output space. Whilst modern applications of multi-output prediction deal with increasingly large data sets, they also include problems where the target dimensionality is increasingly large. For example, the number of categories in multi-label learning is often of the order of tens of thousands, an emergent problem referred to as *extreme classification* (Agrawal et al., 2013; Babbar & Schölkopf, 2017; Bhatia et al., 2015; Jain et al., 2019).

Formally, the task of multi-output prediction is to learn a vector-valued function from a labelled training set. A common tool in the theoretical analysis of this problem has been a vector-valued extension of Talagrand's contraction inequality for Lipschitz losses (Ledoux & Talagrand, 2013). Both (Maurer, 2016) and (Cortes et al., 2016) established vector-contraction inequalities for Rademacher complexity that gave rise to learning guarantees for multi-output prediction problems with a linear dependence on the dimensionality of the output space. More recently, (Lei et al., 2019) has provided more refined vector-contraction inequalities for both Gaussian and Rademacher complexity. This approach leads to a highly favourable sub-linear dependence on the output dimensionality, which can even be logarithmic, depending on the degree of regularisation. These structural results lead to a slow convergence rate $O(n^{-1/2})$. Guermeur (2017) and Musayeva et al. (2019) explore an alternative approach based on covering numbers. (Chzhen et al., 2017) derived a bound for multi-label classification based on Rademacher complexities. Each of these bounds give rise to favourable dependence on the dimensionality of the output space, but with a slow rate of order $O(n^{-1/2})$.

Local Rademacher complexities provide a crucial tool in establishing faster rates of convergence (Bousquet, 2002; Bartlett et al., 2005; Koltchinskii et al., 2006; Lei et al., 2016). By leveraging local Rademacher complexities, Liu et al. (2019) have derived guarantees for multi-class learning with function classes that are linear in an RKHS, building on their previous margin based guarantees (Lei et al., 2015; Li et al., 2019). This gives rise to fast rates under suitable spectral conditions. Fast rates of convergence have also been derived by Xu et al. (2016) for multi-label classification with linear function spaces. On the other hand, Chzhen (2019) have derived fast rates of convergence by exploiting an analogue of the margin assumption.

[1]School of Computer Science, University of Birmingham. Correspondence to: Henry W.J. Reeve <henrywjreeve@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

In this paper we establish generalisation bounds for multi-output prediction, which yields fast rates whenever the empirical error is small. We address this problem by generalising to vector-valued functions a smoothness based approach due to (Srebro et al., 2010). A key advantage of our approach is that it allow us to accommodate a wide variety of multi-output loss functions, and hypothesis classes, making our analytic framework applicable to a variety of learning tasks. Below we summarise the contributions of this paper:

- We give a contraction inequality for the local Rademacher complexity of vector-valued functions (Proposition 1). The main ingredient is a self-bounding Lipschitz condition for multi-output loss functions that holds for several widely used examples.
- We leverage our localised contraction inequality to give a general upper bound for multi-output learning (Theorem 1), which exhibits fast rates whenever the empirical error is small.
- We demonstrate a concrete use our general result, by derive from it a state-of-the-art bound for ensembles of multi-output decision trees (Theorem 7).

Furthermore, the obtained rates on multi-output learning are minimax optimal up to logarithmic factors. The corresponding lower bounds can be found in the full version (Reeve & Kabán, 2020).

1.1. Problem Setting

We shall consider multi-output prediction problems in supervised learning. Suppose we have a measurable space $\mathcal{X}$, a label space $\mathcal{Y}$ and an output space $\mathcal{V}$. We shall assume that there is an unknown probability distribution P over random variables (X, Y), taking values in $\mathcal{X} \times \mathcal{Y}$. The performance is quantified through a loss function $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$.

Let $\mathcal{M}(\mathcal{X}, \mathcal{V})$ denote the set of measurable functions $f : \mathcal{X} \to \mathcal{V}$. The goal of the learner is to obtain $f \in \mathcal{M}(\mathcal{X}, \mathcal{V})$ such that the corresponding risk $\mathcal{E}_{\mathcal{L}}(f, P) := \mathbb{E}_{(X,Y)\sim P}[\mathcal{L}(f(X), Y)]$ is as low as possible. The learner selects $f \in \mathcal{M}(\mathcal{X}, \mathcal{V})$ based upon a sample $\mathcal{D} := \{(X_i, Y_i)\}_{i\in[n]}$, where (X_i, Y_i) are independent copies of (X, Y). We let $\hat{\mathcal{E}}_{\mathcal{L}}(f, \mathcal{D}) := n^{-1} \cdot \sum_{i\in[n]} \mathcal{L}(f(X_i), Y_i)$ denote the empirical risk. When the distribution P and the sample $\mathcal{D}$ are clear from context we shall write $\mathcal{E}_{\mathcal{L}}(f)$ in place of $\mathcal{E}_{\mathcal{L}}(f, P)$ and $\hat{\mathcal{E}}_{\mathcal{L}}(f)$ in place of $\hat{\mathcal{E}}_{\mathcal{L}}(f, \mathcal{D})$. We consider *multi-output* prediction problems in which $\mathcal{V} \subseteq \mathbb{R}^q$. We let $\|\cdot\|_\infty$ denote the max norm on $\mathbb{R}^q$ and for a positive integer $m \in \mathbb{N}$ we let $[m] := \{1, \cdots, m\}$.

2. The Self-bounding Lipschitz Condition

We introduce the following *self-bounding Lipschitz* condition for multi-output loss functions.

Definition 1 (Self-bounding Lipschitz condition)**.** *A loss function $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ is said to be* (λ, θ)-self-bounding Lipschitz *for $\lambda, \theta \geq 0$ if for all $y \in \mathcal{Y}$ and $u, v \in \mathcal{V}$,*

$$|\mathcal{L}(u, y) - \mathcal{L}(v, y)| \leq \lambda \cdot \max\{\mathcal{L}(u, y), \mathcal{L}(v, y)\}^\theta \cdot \|u - v\|_\infty.$$

This condition interpolates continuously between a classical Lipschitz condition (when $\theta = 0$) and a multi-dimensional analogue of a smoothness condition (when $\theta = 1/2$), and will be the main assumption that we use to obtain our results.

Our motivation for introducing Definition 1 is as follows. Firstly, in recent work of (Lei et al., 2019) the classical Lipschitz condition with respect to the ℓ_∞ norm has been utilised to derive multi-class bounds with a favourable dependence upon the number of classes q. The role of the ℓ_∞ norm is crucial since it prevents the deviations in the loss function from accumulating as the output dimension q grows. Our goal is to give a general framework which simultaneously achieves a favourable dependence upon n. Secondly, Srebro et al. (2010) introduced a second-order smoothness condition on the loss function. This condition corresponds to the special case whereby $q = 1$ and $\theta = 1/2$. Srebro et al. (2010) showed that this smoothness condition gives rise to an optimistic bound having a fast rate $O(n^{-1})$ in the realisable case. The self-bounding Lipschitz condition provides a multi-dimensional analogue of this condition when $\theta = 1/2$, intended to yield a favourable dependence on the number of samples n. The results established in Sections 3 and 5 show that this is indeed the case, while we also obtain favourable dependence on the number of classes q. Finally, by considering the range of exponents $\theta \in [0, 1/2]$ we exhibit convergence rates ranging from slow $O(n^{-1/2})$ to fast $O(n^{-1})$ in the realisable case. This is reminiscent of the celebrated Tsybakov margin condition (Mammen & Tsybakov, 1999), which interpolates between slow and fast rates in the parametric classification setting. Crucially, however, whilst the Tsybakov margin condition (Mammen & Tsybakov, 1999) is a condition on the underlying distribution – which cannot be verified in practice – the self-bounding Lipschitz condition is a property of a loss function – which may be verified analytically by the learner.

2.1. Verifying the Self-bounding Lipschitz Condition

We start by giving a collection of results which can be used to verify that a given loss function satisfies the self-bounding Lipschitz condition. The following lemma is proved in the Supplementary Appendix A.

Lemma 1. *Take any $\lambda > 0$, $\theta \in [0, 1/2]$. Suppose that $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0, \infty)$ is a loss function such that for any $u \in \mathcal{V}$, $y \in \mathcal{Y}$, there exists a non-negative differentiable function $\varphi_{u,y} : \mathbb{R} \to [0, \infty)$ satisfying*

1. *$\varphi_{u,y}(0) = \mathcal{L}(u, y)$;*

2. $\forall t > 0$, $\sup_{v:\|u-v\|_\infty \le t}\{\mathcal{L}(v,y)\} \le \varphi_{u,y}(t)$.
3. *The derivative* $\varphi'_{u,y}(t)$ *is non-negative on* $[0,\infty)$*;*
4. $\forall t_0, t_1 \in \mathbb{R}$, $|\varphi'_{u,y}(t_1) - \varphi'_{u,y}(t_0)| \le \left(\frac{\lambda}{2}\right)^{\frac{1}{1-\theta}} \cdot |t_1 - t_0|^{\frac{\theta}{1-\theta}}$*;*

Then $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0,\infty)$ *is* (λ,θ)*-self-bounding Lipschitz.*

The following Lemma shows that clipping preserves this condition.

Lemma 2. *Suppose that* $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0,\infty)$ *is a* (λ,θ)*-self-bounding Lipschitz loss function with* $\lambda > 0$, $\theta \in [0,1]$*. Then the loss* $\tilde{\mathcal{L}} : \mathcal{V} \times \mathcal{Y} \to [0,b]$ *defined by* $\tilde{\mathcal{L}}(u,y) = \min\{\mathcal{L}(u,y), b\}$ *is* (λ,θ)*-self-bounding Lipschitz.*

Finally, we note the following monotonicity property, which follows straightforwardly from the definition.

Lemma 3. *Suppose that* $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0,b]$ *is a bounded* (λ,θ)*-self-bounding Lipschitz loss function with* $\lambda > 0$, $\theta \in [0,1]$*. Then given any* $\tilde{\theta} \le \theta$*, the loss* $\mathcal{L}$ *is also* $(\tilde{\lambda},\tilde{\theta})$*-self-bounding Lipschitz with* $\tilde{\lambda} = \lambda \cdot b^{\theta-\tilde{\theta}}$*.*

2.2. Examples

We now demonstrate several examples of multi-output loss functions that satisfy our self-bounding Lipschitz condition. In each of the examples below we shall show that the self-bounding Lipschitz condition is satisfied by applying our sufficient condition (Lemma 1). Detailed proofs are given in the Supplementary Appendix A.

2.2.1. Multi-class Losses

We begin with the canonical multi-output prediction problem of multi-class classification in which $\mathcal{Y} = [q]$ and $\mathcal{V} = \mathbb{R}^q$. A popular loss function for the theoretical analysis of multi-class learning is the margin loss (Crammer & Singer, 2001). The smoothed analogue of the margin loss was introduced by Srebro et al. (2010) in the one-dimensional setting, and Li et al. (2018) in the multi-class setting.

Example 1 (Smooth margin losses)**.** *Given* $\mathcal{Y} = [q]$ *we define the margin function* $m : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ *by* $m(u,y) := u_y - \max_{j\in[q]\setminus\{y\}}\{u_j\}$*. The* zero-one loss $\mathcal{L}_{0,1} : \mathcal{V} \times \mathcal{Y} \to [0,1]$ *is defined by* $\mathcal{L}_{0,1}(u,y) = \mathbf{1}\{m(u,y) \le 0\}$*. Whilst natural, the zero-one loss has the drawback of being discontinuous, which presents an obstacle for deriving guarantees. For each* $\rho > 0$*, the corresponding margin loss* $\mathcal{L}_\rho : \mathcal{V} \times \mathcal{Y} \to [0,1]$ *is defined by* $\mathcal{L}_\rho(u,y) = \mathbf{1}\{m(u,y) \le \rho\}$*. The margin loss* $\mathcal{L}_\rho$ *is also discontinuous. However, we may define a* smooth margin loss $\tilde{\mathcal{L}}_\rho : \mathcal{V} \times \mathcal{Y} \to [0,1]$ *by* $\tilde{\mathcal{L}}_\rho(u,y)$

$$:= \begin{cases} 1 & \text{if } m(u,y) \le 0 \\ 2\left(\frac{m(u,y)}{\rho}\right)^3 - 3\left(\frac{m(u,y)}{\rho}\right)^2 + 1 & \text{if } m(u,y) \in [0,\rho] \\ 0 & \text{if } m(u,y) \ge \rho. \end{cases}$$

By applying Lemma 1 we can show that $\tilde{\mathcal{L}}_\rho$ *is* (λ,θ)-self-bounding Lipschitz *with* $\lambda = 4\sqrt{6} \cdot \rho^{-1}$ *and* $\theta = 1/2$*. Moreover, the smooth margin loss satisfies* $\mathcal{L}_{0,1}(u,y) \le \tilde{\mathcal{L}}_\rho(u,y) \le \mathcal{L}_\rho(u,y)$ *for* $(u,y) \in \mathcal{V} \times \mathcal{Y}$*.*

The margin loss plays a central role in learning theory and continues to receive significant attention in the analysis of multi-class prediction (Guermeur, 2017; Li et al., 2018; Musayeva et al., 2019), so it is fortuitous that our self-bounding Lipschitz condition incorporates the smooth margin loss. More importantly, however, the self-bounding Lipschitz condition applies to a variety of other loss functions which have received less attention in statistical learning theory.

One of the most widely used loss functions in practical applications is the *multinomial logistic loss*, also known as the softmax loss.

Example 2 (Multinomial logistic loss)**.** *Given* $\mathcal{Y} = [q]$*, the* multinomial logistic loss $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0,\infty)$ *is defined by*

$$\mathcal{L}(u,y) = \log\left(\sum_{j\in[q]} \exp(u_j - u_y)\right),$$

where $u = (u_j)_{j\in[q]}$ *and* $y \in [q]$*. For each* $(u,y) \in \mathcal{V} \times [q]$ *let* $A_{u,y} = \sum_{j\in[q]\setminus\{y\}} \exp(u_j - u_y)$ *and define* $\varphi_{u,y}(t) = \log\left(1 + A_{u,y} \cdot \exp(2t)\right)$*. By applying Lemma 1 with* $\varphi_{u,y}$ *we can show that the multinomial logistic loss* $\mathcal{L}$ *is* (λ,θ)-self-bounding Lipschitz *with* $\lambda = 2$ *and* $\theta = 1/2$*.*

Recently, Lei et al. (2019) pointed out that the multinomial-logistic loss is 2-Lipschitz with respect to the ℓ_∞-norm (equivalently, $(2,0)$-self-bounding Lipschitz). This gives rise to a slow rate of order $O(n^{-1/2})$. The fact that the multinomial-logistic loss is also $(2,1/2)$-self bounding can be used to derive more favourable guarantees, as we shall see in Section 3.

2.2.2. Multi-label Losses

In multi-label prediction instances may be simultaneously assigned to several categories. We have $\mathcal{Y} \subseteq \{0,1\}^q$, where q is the total number possible classes. Whilst q is often very large, the total number of simultaneous labels is typically much smaller. Hence, we consider the set of k-sparse binary vectors $\mathbb{S}(k) = \{(y_j)_{j\in[q]} \in \{0,1\}^q : \sum_{j\in[q]} y_j \le k\}$ denote the set of k-sparse vectors, where $k \le [q]$. We consider the pick-all-labels loss (Menon et al., 2019; Reddi et al., 2019).

Example 3 (Pick-all-labels)**.** *Given* $\mathcal{Y} = \mathbb{S}(k)$*, the* pick-all-labels loss $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0,\infty)$ *is defined by*

$$\mathcal{L}(u,y) = \textstyle\sum_{l\in[q]} y_l \log\left(\sum_{j\in[q]} \exp(u_j - u_l)\right),$$

where $u = (u_j)_{j\in[q]} \in \mathcal{V}$ *and* $y = (y_j)_{j\in[q]} \in \mathcal{Y}$*. For each* $(u,y) \in \mathcal{V} \times \mathcal{Y}$ *we define* $\varphi_{u,y} : \mathbb{R} \to [0,\infty)$

by $A_{u,y} = \sum_{j\in[q]\setminus\{l\}} \exp(u_j - u_l)$ *and let* $\varphi_{u,y}(t) := \sum_{l\in[q]} y_l \log\left(1 + A_{u,y}\cdot\exp(2t)\right)$. *By applying Lemma 1 with* $\varphi_{u,y}$ *we can show that* $\mathcal{L}$ *is* (λ,θ)-self-bounding Lipschitz *with* $\lambda = 2\sqrt{k}$ *and* $\theta = 1/2$.

Crucially, the constant λ for the pick-all-labels family of losses is a function of the sparsity k, rather than the total number of labels. This means that our approach is applicable to multi-label problems with with tens of thousands of labels, as long as the label-vectors are k-sparse.

2.2.3. Losses for Multi-target Regression

We now return to the problem of *multi-target regression* in which $\mathcal{Y} = \mathbb{R}^q$ (Borchani et al., 2015).

Example 4 (Sup-norm losses)**.** *Given* $\kappa,\ \gamma \in [1,2]$ *we can define a loss-function* $\mathcal{L} : \mathcal{V}\times\mathcal{Y} \to \mathbb{R}$ *for* multi-target regression *by setting* $\mathcal{L}(u,y) = \kappa\cdot\|u-y\|_\infty^\gamma$. *By applying Lemma 1 with* $\varphi_{u,y}(t) = \kappa\cdot(\|u-y\|_\infty + t)^\gamma$ *we can see that* $\mathcal{L}$ *is a* (λ,θ)-self-bounding Lipschitz *with* $\lambda = (8\kappa)^{1-\theta}$ *and* $\theta = (\gamma-1)/\gamma$. *This yields examples of* (λ,θ)-self-bounding Lipschitz *loss functions for all* $\lambda > 0$ *and* $\theta \in [0,1/2]$.

With these examples in mind we are ready to present our results.

3. Main Results

In this section we give a general upper bound for multi-output prediction problems under the self-bounding Lipschitz condition. A key tool for proving this result will be a contraction inequality for local Rademacher complexity of vector valued functions given in Section 4.1, and which may also be of independent interest. First, we recall the concept of Rademacher complexity.

Definition 2 (Rademacher complexity)**.** *Let* $\mathcal{Z}$ *be a measurable space and consider a function class* $\mathcal{G}\subseteq\mathcal{M}(\mathcal{Z},\mathbb{R})$. *Given a sequence* $\boldsymbol{z} = (z_i)\in\mathcal{Z}^n$ *we define the* empirical Rademacher complexity *of* $\mathcal{G}$ *with respect to* $\boldsymbol{z}$ *by*[1]

$$\hat{\mathfrak{R}}_{\boldsymbol{z}}(\mathcal{G}) := \sup_{\tilde{\mathcal{G}}\subseteq\mathcal{G}:|\tilde{\mathcal{G}}|<\infty} \mathbb{E}_{\boldsymbol{\sigma}}\left(\sup_{g\in\tilde{\mathcal{G}}}\frac{1}{n}\sum_{i\in[n]}\sigma_i\cdot g(z_i)\right),$$

where the expectation is taken over sequences of independent Rademacher random variables $\boldsymbol{\sigma} = (\sigma_i)_{i\in[n]}$ *with* $\sigma_i \in \{-1,+1\}^n$. *For each* $n\in\mathbb{N}$, *the* worst-case Rademacher complexity *of* $\mathcal{G}$ *is defined by* $\mathfrak{R}_n(\mathcal{G}) := \sup_{\boldsymbol{z}\in\mathcal{Z}^n}\hat{\mathfrak{R}}_{\boldsymbol{z}}(\mathcal{G})$.

The Rademacher complexity is defined in the context of real-valued functions. However, in this work we deal with multi-output prediction so we shall focus on function classes $\mathcal{F}\subseteq\mathcal{M}(\mathcal{X},\mathbb{R}^q)$. In order to utilise the theory of Rademacher complexity in this context we shall transform function classes $\mathcal{F}\subseteq\mathcal{M}(\mathcal{X},\mathbb{R}^q)$ into the projected function classes $\Pi\circ\mathcal{F}\subseteq\mathcal{M}(\mathcal{X}\times[q],\mathbb{R})$ as follows. Firstly, for each $j\in[q]$ we define $\pi_j : \mathbb{R}^q\to\mathbb{R}$ to be the projection onto the j-th coordinate. We then define, for each $f\in\mathcal{M}(\mathcal{X},\mathbb{R}^q)$, the function $\Pi\circ f : \mathcal{X}\times[q]\to\mathbb{R}$ by $(\Pi\circ f)(x,j) = \pi_j(f(x))$. Finally, given $\mathcal{F}\subseteq\mathcal{M}(\mathcal{X},\mathbb{R}^q)$ we let $\Pi\circ\mathcal{F} := \{\Pi\circ f : f\in\mathcal{F}\}\subseteq\mathcal{M}(\mathcal{X}\times[q],\mathbb{R})$.

Our central result is the following relative bound.

Theorem 1. *Suppose we have a class of multi-output functions* $\mathcal{F}\subseteq\mathcal{M}(\mathcal{X},[-\beta,\beta]^q)$, *and a* (λ,θ)*-self-bounding Lipschitz loss function* $\mathcal{L}:\mathcal{V}\times\mathcal{Y}\to[0,b]$ *for some* $\beta,b\geq 1$, $\lambda>0$, $\theta\in[0,1/2]$. *Take* $\delta\in(0,1)$, $n\in\mathbb{N}$ *and let*

$$\Gamma^{\lambda,\theta}_{n,q,\delta}(\mathcal{F}) := \left(\lambda\left(\sqrt{q}\cdot\log^{3/2}(e\beta nq)\cdot\mathfrak{R}_{nq}(\Pi\circ\mathcal{F}) + \tfrac{1}{\sqrt{n}}\right)\right)^{\frac{1}{1-\theta}} + \tfrac{b}{n}\cdot(\log(1/\delta) + \log(\log n)).$$

There exists numerical constants $C_0, C_1 > 0$ *such that given an i.i.d. sample* $\mathcal{D}$ *the following holds with probability at least* $1-\delta$ *for all* $f\in\mathcal{F}$,

$$\mathcal{E}_{\mathcal{L}}(f) \leq \hat{\mathcal{E}}_{\mathcal{L}}(f) + C_0\cdot\left(\sqrt{\hat{\mathcal{E}}_{\mathcal{L}}(f)\cdot\Gamma^{\lambda,\theta}_{n,q,\delta}(\mathcal{F})} + \Gamma^{\lambda,\theta}_{n,q,\delta}(\mathcal{F})\right).$$

Moreover, if $f^*\in\operatorname{argmin}_{f\in\mathcal{F}}\{\mathcal{E}_{\mathcal{L}}(f)\}$ *minimises the risk and* $\hat{f}\in\operatorname{argmin}_{f\in\mathcal{F}}\{\hat{\mathcal{E}}_{\mathcal{L}}(f)\}$ *minimises the empirical risk, then with probability at least* $1-\delta$,

$$\mathcal{E}_{\mathcal{L}}(\hat{f}) \leq \mathcal{E}_{\mathcal{L}}(f^*) + C_1\cdot\left(\sqrt{\mathcal{E}_{\mathcal{L}}(f^*)\cdot\Gamma^{\lambda,\theta}_{n,q,\delta}(\mathcal{F})} + \Gamma^{\lambda,\theta}_{n,q,\delta}(\mathcal{F})\right).$$

The proof of Theorem 1 is given in Section 4.2. It builds on a local contraction inequality result (Proposition 1, Section 4.1), combined with techniques from (Bousquet, 2002).

Theorem 1 gives an upper bound for the generalisation gap $(\mathcal{E}_{\mathcal{L}}(f) - \hat{\mathcal{E}}_{\mathcal{L}}(f))$, framed in terms of a complexity term $\Gamma^{\lambda,\theta}_{n,q,\delta}(\mathcal{F})$, which depends upon both the Rademacher complexity of the projected function class $\mathfrak{R}_{nq}(\Pi\circ\mathcal{F})$ and the self-bounding Lipschitz parameters λ, θ. When the empirical error is small in relation to the complexity term $(\hat{\mathcal{E}}_{\mathcal{L}}(f) \leq \Gamma^{\lambda,\theta}_{n,q,\delta}(\mathcal{F}))$, the generalisation gap is of order $\Gamma^{\lambda,\theta}_{n,q,\delta}(\mathcal{F})$. In less favourable circumstances we recover a bound of order $\sqrt{\Gamma^{\lambda,\theta}_{n,q,\delta}(\mathcal{F})}$.

3.1. Comparison with State of the Art

In this section we compare our main result (Theorem 1) with a closely related guarantee due to Lei et al. (2019). Observe that a loss function $\mathcal{L}$ is λ-Lipschitz if it is (λ,θ)-self-bounding Lipschitz with $\theta = 0$.

Theorem 2. *(Lei et al., 2019) Suppose we have a class of multi-output functions* $\mathcal{F}\subseteq\mathcal{M}(\mathcal{X},[-\beta,\beta]^q)$, *and a* λ-

[1] Taking the supremum over finite subsets $\tilde{\mathcal{G}}\subseteq\mathcal{G}$ is required to ensure that the function within the expectation is measurable (Talagrand, 2014). This technicality can typically be overlooked.

Lipschitz loss function $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0, b]$ *for some* $\beta, b \geq 1$ *and* $\lambda > 0$. *Take* $\delta \in (0, 1)$, $n \in \mathbb{N}$ *and let*

$$\mathfrak{J}^{\lambda}_{n,q,\delta}(\mathcal{F}) := \lambda \left(\sqrt{q} \cdot \log^{3/2}(e\beta nq) \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F}) + \tfrac{1}{\sqrt{n}} \right).$$

There exists numerical constants $C_2, C_3 > 0$ *such that given an i.i.d. sample* $\mathcal{D}$ *the following holds with probability at least* $1 - \delta$ *for all* $f \in \mathcal{F}$,

$$\mathcal{E}_{\mathcal{L}}(f) \leq \hat{\mathcal{E}}_{\mathcal{L}}(f) + C_2 \cdot \mathfrak{J}^{\lambda}_{n,q,\delta}(\mathcal{F}) + b\sqrt{\frac{\log(1/\delta)}{n}}.$$

Moreover, if $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} \{\mathcal{E}_{\mathcal{L}}(f)\}$ *minimises the risk and* $\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} \{\hat{\mathcal{E}}_{\mathcal{L}}(f)\}$ *minimises the empirical risk, then with probability at least* $1 - \delta$,

$$\mathcal{E}_{\mathcal{L}}(\hat{f}) \leq \mathcal{E}_{\mathcal{L}}(f^*) + C_3 \cdot \mathfrak{J}^{\lambda}_{n,q,\delta}(\mathcal{F}) + 2b\sqrt{\frac{\log(1/\delta)}{n}}.$$

Theorem 2 is a mild generalisation of Theorem 6 from (Lei et al., 2019), originally formulated for multi-class classification and $\mathcal{F}$ an RKHS. For completeness we show that Theorem 2 follows from Proposition 1 in the Supplementary Appendix B. . Note that by the monotonicity property (Lemma 3) any loss function $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0, b]$ which is (λ, θ)-self-bounding Lipschitz is also $\lambda \cdot b^{\theta}$-Lipschitz, so the additive bound in Theorem 2 also applies.

To gain a deeper intuition for the bound in Theorem 1 we compare with the bound in Theorem 2. Let's suppose that $\mathfrak{R}_{nq}(\Pi \circ \mathcal{F}) = \tilde{O}((nq)^{-1/2})$ (for a concrete example where this is the case see Section 5). We then have $\Gamma^{\lambda,\theta}_{n,q,\delta}(\mathcal{F}) = \tilde{O}(n^{-\frac{1}{2(1-\theta)}})$. For large values of $\hat{\mathcal{E}}_{\mathcal{L}}(f)$ Theorem 1 gives a bound on generalisation gap $(\mathcal{E}_{\mathcal{L}}(f) - \hat{\mathcal{E}}_{\mathcal{L}}(f))$ of order $\tilde{O}(n^{-\frac{1}{4(1-\theta)}})$, which is slower than the rate achieved by Theorem 2 whenever $\theta < 1/2$. However, when $\hat{\mathcal{E}}_{\mathcal{L}}(f)$ is small ($\hat{\mathcal{E}}_{\mathcal{L}}(f) \leq \tilde{O}(n^{-\frac{1}{2(1-\theta)}})$), Theorem 1 gives rise to a bound of order $\tilde{O}(n^{-\frac{1}{2(1-\theta)}})$, yielding faster rates than can be obtained through the standard Lipschitz condition alone whenever $\theta > 0$. Finally note that if the loss $\mathcal{L}$ is (λ, θ)-self-bounding Lipschitz with $\theta = 1/2$ then the rates given by Theorem 1 always either match or outperform the rates given by Theorem 2. Moreover, $\theta = 1/2$ occurs for several practical examples discussed in Section 2.2 including the multinomial-logistic loss.

4. Proofs of Main Results

We now turn to stating and proving the key ingredient of our main result, Proposition 1.

4.1. A Contraction Inequality for the Local Rademacher Complexity of Vector-valued Function Classes

First we introduce some additional notation. Suppose $f \in \mathcal{M}(\mathcal{X}, \mathcal{V})$. Given a loss function $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ we define $\mathcal{L} \circ f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ by $(\mathcal{L} \circ f)(x, y) = \mathcal{L}(f(x), y)$. We extend this definition to function classes $\mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{V})$ by $\mathcal{L} \circ \mathcal{F} = \{\mathcal{L} \circ f : f \in \mathcal{F}\}$. Moreover, for each $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ and $r > 0$, a subset $\mathcal{F}|^r_{\boldsymbol{z}} := \{f \in \mathcal{F} : \ \hat{\mathcal{E}}_{\mathcal{L}}(f, \boldsymbol{z}) \leq r\}$. Intuitively, the local Rademacher complexity allows us to *zoom in* upon the neighbourhood of the empirical risk minimiser. This is the subset that matters in practice and is typically much smaller than the full $\Pi \circ \mathcal{F}$.

Proposition 1. *Suppose we have a class of multi-output functions* $\mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, [-\beta, \beta]^q)$, *where* $\beta \geq 1$. *Given a* (λ, θ)*-self-bounding Lipschitz loss function* $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0, \mathbb{R}]$, *where* $\lambda > 0$, $\theta \in [0, 1/2]$ *and* $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, $r > 0$, *we have the following bound,*

$$\hat{\mathfrak{R}}_{\boldsymbol{z}}\left(\mathcal{L} \circ \mathcal{F}|^r_{\boldsymbol{z}}\right) \leq \lambda r^{\theta} \left(2^9 \sqrt{q} \cdot \log^{3/2}(e\beta nq) \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F}) + n^{-1/2} \right).$$

The proof of Proposition 1, given later in this section, relies upon covering numbers.

Definition 3 (Covering numbers)**.** *Let* $(\mathcal{M}, \rho)$ *be a semi-metric space. Given a set* $A \subseteq \mathcal{M}$ *and an* $\epsilon > 0$, *a subset* $\tilde{A} \subseteq A$ *is said to be a (proper)* ϵ*-cover of* A *if, for all* $a \in A$, *there exists some* $\tilde{a} \in \tilde{A}$ *with* $\rho(a, \tilde{a}) \leq \epsilon$. *We let* $\mathcal{N}(\epsilon, A, \rho)$ *denote the minimal cardinality of an* ϵ*-cover for* A.

We shall consider covering numbers for two classes of data-dependent semi-metric spaces. Let $\mathcal{Z}$ be a measurable space and take $\mathcal{G} \subseteq \mathcal{M}(\mathcal{Z}, \mathbb{R})$. For each $n \in \mathbb{N}$ and each sequence $\boldsymbol{z} = (z_i)_{i \in [n]} \in \mathcal{Z}^n$ we define a pair of metrics $\rho_{\boldsymbol{z},2}$ and $\rho_{\boldsymbol{z},\infty}$ by

$$\rho_{\boldsymbol{z},2}(g_0, g_1) := \sqrt{\frac{1}{n} \sum_{i \in [n]} (g_0(z_i) - g_1(z_i))^2}$$

$$\rho_{\boldsymbol{z},\infty}(g_0, g_1) := \max_{i \in [n]} \{|g_0(z_i) - g_1(z_i)|\},$$

where $g_0, g_1 \in \mathcal{G}$. The first stage of the proof of Proposition 1 will be using the following lemma which bounds the covering number of $\mathcal{L} \circ \mathcal{F}|^r_{\boldsymbol{z}}$ in terms of an associated covering number for $\Pi(\mathcal{F})$.

Lemma 4. *Suppose that* $\mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, \mathbb{R}^q)$ *and* $\mathcal{L}$ *is* (λ, θ)*-self-bounding Lipschitz with* $\theta \in [0, 1/2]$. *Take* $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0, b]$, $\boldsymbol{z} = \{(x_i, y_i)\}_{i \in [n]} \in (\mathcal{X} \times \mathcal{Y})^n$, $r > 0$ *and define* $\boldsymbol{w} = \{(x_i, j)\}_{(i,j) \in [n] \times [q]} \in (\mathcal{X} \times [q])^{nq}$. *Given any* $f_0, f_1 \in \mathcal{F}|^r_{\boldsymbol{z}}$ *we have*

$$\rho_{\boldsymbol{z},2}(\mathcal{L} \circ f_0, \mathcal{L} \circ f_1) \leq 2^{\theta} \lambda r^{\theta} \cdot \rho_{\boldsymbol{w},\infty}(\Pi \circ f_0, \Pi \circ f_1).$$

Moreover, for any $\epsilon > 0$,

$$\mathcal{N}\left(2^{1+\theta}\lambda r^\theta \cdot \epsilon, \mathcal{L}\circ\mathcal{F}|^r_{\boldsymbol{z}}, \rho_{\boldsymbol{z},2}\right) \leq \mathcal{N}\left(\epsilon, \Pi\circ\mathcal{F}, \rho_{\boldsymbol{w},\infty}\right).$$

Proof of Lemma 4. To prove the first part of the lemma we take $f_0, f_1 \in \mathcal{F}|^r_{\boldsymbol{z}}$ and let $\zeta = \rho_{\boldsymbol{w},\infty}(\Pi\circ f_0, \Pi\circ f_1)$. It follows from the construction of $\boldsymbol{w}$ that $|\pi_j(f_0(x_i)) - \pi_j(f_1(x_i))| \leq \zeta$ for each $(i,j) \in [n]\times[q]$, so $\|f_0(x_i) - f_1(x_i)\|_\infty \leq \zeta$ for each $i \in [n]$.

Furthermore, by the self-bounding Lipschitz condition we deduce that for each $i \in [n]$,

$$\begin{aligned}
&|\mathcal{L}(f_0(x_i), y_i) - \mathcal{L}(f_1(x_i), y_i)| \\
&\leq \lambda \cdot \max\left\{\mathcal{L}(f_0(x_i), y_i), \mathcal{L}(f_1(x_i), y_i)\right\}^\theta \\
&\qquad\qquad \cdot \|f_0(x_i) - f_1(x_i)\|_\infty \\
&\leq \lambda \cdot \max\left\{\mathcal{L}(f_0(x_i), y_i), \mathcal{L}(f_1(x_i), y_i)\right\}^\theta \cdot \zeta.
\end{aligned}$$

Hence, by Jensen's inequality we have

$$\begin{aligned}
&\rho_{\boldsymbol{z},2}(\mathcal{L}\circ f_0, \mathcal{L}\circ f_1)^2 \\
&= \frac{1}{n}\sum_{i\in[n]} \left(\mathcal{L}(f_0(x_i), y_i) - \mathcal{L}(f_1(x_i), y_i)\right)^2 \\
&\leq (\lambda\zeta)^2 \cdot \frac{1}{n}\sum_{i\in[n]} \max\left\{\mathcal{L}(f_0(x_i), y_i), \mathcal{L}(f_1(x_i), y_i)\right\}^{2\theta} \\
&\leq (\lambda\zeta)^2 \cdot \left(\frac{1}{n}\sum_{i\in[n]} \max\left\{\mathcal{L}(f_0(x_i), y_i), \mathcal{L}(f_1(x_i), y_i)\right\}\right)^{2\theta} \\
&\leq (\lambda\zeta)^2 \cdot \left(\hat{\mathcal{E}}_{\mathcal{L}}(f_0, \boldsymbol{z}) + \hat{\mathcal{E}}_{\mathcal{L}}(f_1, \boldsymbol{z})\right)^{2\theta} \leq (\lambda\zeta)^2 \cdot (2r)^{2\theta},
\end{aligned}$$

where we use the fact that $\theta \in [0, 1/2]$ and $\max\{\hat{\mathcal{E}}_{\mathcal{L}}(f_0, \boldsymbol{z}), \hat{\mathcal{E}}_{\mathcal{L}}(f_1, \boldsymbol{z})\} \leq r$.

Thus, $\rho_{\boldsymbol{z},2}(\mathcal{L}\circ f_0, \mathcal{L}\circ f_1)$

$$\leq 2^\theta \lambda r^\theta \cdot \zeta = 2^\theta \lambda r^\theta \cdot \rho_{\boldsymbol{w},\infty}(\Pi\circ f_0, \Pi\circ f_1).$$

This completes the proof of the first part of the lemma.

To prove the second part of the lemma we note that since $\Pi\circ\mathcal{F}|^r_{\boldsymbol{z}} \subseteq \Pi\circ\mathcal{F}$ we have[2]

$$\mathcal{N}\left(2\epsilon, \Pi\circ\mathcal{F}|^r_{\boldsymbol{z}}, \rho_{\boldsymbol{w},\infty}\right) \leq \mathcal{N}\left(\epsilon, \Pi\circ\mathcal{F}, \rho_{\boldsymbol{w},\infty}\right),$$

so we may choose $f_1, \cdots, f_m \in \mathcal{F}|^r_{\boldsymbol{z}}$ with $m \leq \mathcal{N}\left(\epsilon, \Pi\circ\mathcal{F}, \rho_{\boldsymbol{w},\infty}\right)$ such that $\Pi\circ f_1, \cdots, \Pi\circ f_m$ forms a 2ϵ-cover of $\Pi\circ\mathcal{F}|^r_{\boldsymbol{z}}$ with respect to the $\rho_{\boldsymbol{w},\infty}$ metric.

To complete the proof it suffices to show that $\mathcal{L}\circ f_1, \cdots, \mathcal{L}\circ f_m$ is a $2^{1+\theta}\lambda r^\theta \cdot \epsilon$-cover of $\mathcal{L}\circ\mathcal{F}|^r_{\boldsymbol{z}}$ with respect to the $\rho_{\boldsymbol{z},2}$ metric.

[2]The factor of 2 is required as we are using *proper* covers, which are subsets of the set being covered (see Definition 3).

Take any $\tilde{g} \in \mathcal{L}\circ\mathcal{F}|^r_{\boldsymbol{z}}$, so $\tilde{g} = \mathcal{L}\circ\tilde{f}$ for some $\tilde{f} \in \mathcal{F}|^r_{\boldsymbol{z}}$. Since $\Pi\circ f_1, \cdots, \Pi\circ f_m$ forms a 2ϵ-cover of $\Pi\circ\mathcal{F}|^r_{\boldsymbol{z}}$ we may choose $l \in [m]$ so that $\rho_{\boldsymbol{w},\infty}(\Pi\circ f_l, \Pi\circ\tilde{f}) \leq 2\epsilon$. By the first part of the lemma we deduce that

$$\rho_{\boldsymbol{z},2}(\mathcal{L}\circ f_l, \tilde{g}) = \rho_{\boldsymbol{z},2}(\mathcal{L}\circ f_l, \mathcal{L}\circ\tilde{f}) \leq 2^{1+\theta}\lambda r^\theta \cdot \epsilon$$

Since this holds for all $\tilde{g} \in \mathcal{L}\circ\mathcal{F}|^r_{\boldsymbol{z}}$, we see that $\mathcal{L}\circ f_1, \cdots, \mathcal{L}\circ f_m$ is a $2^{1+\theta}\lambda r^\theta \cdot \epsilon$-cover of $\mathcal{L}\circ\mathcal{F}|^r_{\boldsymbol{z}}$, which completes the proof of the lemma. □

To prove Proposition 1, we shall also utilise two technical results to move from covering numbers to Rademacher complexity and back. First, we shall use the following powerful result from (Srebro et al., 2010) which gives an upper bound for *worst-case* covering numbers in terms of the *worst-case* Rademacher complexity.

Theorem 3 (Srebro et al. (2010))**.** *Given a measurable space $\mathcal{Z}$ and a function class $\mathcal{G} \subseteq \mathcal{M}(\mathcal{Z}, [-\beta, \beta])$, any $\epsilon > 2\cdot\mathfrak{R}_n(\mathcal{G})$ and any $\boldsymbol{z} \in \mathcal{Z}^n$,*

$$\log\mathcal{N}(\epsilon, \mathcal{G}, \rho_{\boldsymbol{z},\infty}) \leq \left(\mathfrak{R}_n(\mathcal{G})\right)^2 \cdot \frac{4n}{\epsilon^2} \cdot \log\frac{2e\beta n}{\epsilon}.$$

We can view this result as an analogue of Sudakov's minoration inequality for ℓ_∞ covers, rather than ℓ_2 covers.

Secondly, we shall use Dudley's inequality (Dudley, 1967) which allows us to bound Rademacher complexities in terms of covering numbers. We shall use the following variant due to (Guermeur, 2017) as it yields more favourable constants.

Theorem 4 (Guermeur (2017))**.** *Suppose we have a measurable space $\mathcal{Z}$, a function class $\mathcal{G} \subseteq \mathcal{M}(\mathcal{Z}, \mathbb{R})$ and a sequence $\boldsymbol{z} \in \mathcal{Z}^n$. For any decreasing sequence $(\epsilon_k)_{k=0}^\infty$ with $\lim_{k\to\infty}\epsilon_k = 0$ with $\epsilon_0 \geq \sup_{g_0,g_1\in\mathcal{G}}\rho_{\boldsymbol{z},2}(g_0, g_1)$, the following inequality holds for all $K \in \mathbb{N}$,*

$$\hat{\mathfrak{R}}_{\boldsymbol{z}}(\mathcal{G}) \leq 2\cdot\sum_{k=1}^{K}(\epsilon_k + \epsilon_{k-1})\cdot\sqrt{\frac{\log\mathcal{N}(\epsilon_k, \mathcal{G}, \rho_{\boldsymbol{z},2})}{n}} + \epsilon_K.$$

We are now ready to complete the proof of our local Rademacher complexity inequality.

Proof of Proposition 1. Take $\boldsymbol{z} = \{(x_i, y_i)\}_{i\in[n]} \in (\mathcal{X}\times\mathcal{Y})^n$ and $r > 0$ and define $\boldsymbol{w} = \{(x_i, j)\}_{(i,j)\in[n]\times[q]} \in (\mathcal{X}\times[q])^{nq}$. By Lemma 4 combined with Theorem 3 applied to $\Pi\circ\mathcal{F}$ we see that for each $\xi > 2\cdot\mathfrak{R}_{nq}(\Pi\circ\mathcal{F})$ we have

$$\begin{aligned}
&\log\mathcal{N}\left(2^{1+\theta}\lambda r^\theta\cdot\xi, \mathcal{L}\circ\mathcal{F}|^r_{\boldsymbol{z}}, \rho_{\boldsymbol{z},2}\right) \\
&\leq \log\mathcal{N}(\xi, \Pi\circ\mathcal{F}, \rho_{\boldsymbol{w},\infty}) \\
&\leq \left(\mathfrak{R}_{nq}(\Pi\circ\mathcal{F})\right)^2 \cdot \frac{4nq}{\xi^2}\cdot\log\frac{2e\beta nq}{\xi}. \qquad (1)
\end{aligned}$$

Moreover, given any $g_0 = \mathcal{L} \circ f_0$, $g_1 = \mathcal{L} \circ f_1 \in \mathcal{L} \circ \mathcal{F}|_{\boldsymbol{z}}^r$, so $\rho_{\boldsymbol{w},\infty}(\Pi \circ f_0, \Pi \circ f_1) \leq 2\beta$, so by the first part of Lemma 4 we have $\rho_{\boldsymbol{z},2}(g_0, g_1) \leq 2^{1+\theta}\lambda r^\theta \cdot \beta$.

Now construct $(\epsilon_k)_{k=0}^\infty$ by $\epsilon_k = 2^{1+\theta}\lambda r^\theta \cdot \beta \cdot 2^{-k}$ and choose

$$K = \lceil \log_2 \left(\beta \cdot \min\{(2 \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F}))^{-1}, (8\sqrt{n})\} \right) \rceil - 1$$

Hence, $\sup_{g_0, g_1 \in \Pi \circ \mathcal{F}|_{\boldsymbol{z}}^r} \rho_{\boldsymbol{z},2}(g_0, g_1) \leq \epsilon_0$ and $\beta \cdot 2^{-K-1} \leq \max\{2 \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F}), (8\sqrt{n})^{-1}\} < \beta \cdot 2^{-K}$.

Furthermore, for $k \leq K$ by letting $\xi_k = \beta \cdot 2^{-k}$, we have $\epsilon_k = 2^{1+\theta}\lambda r^\theta \cdot \xi_k$ and $\xi_k > \max\{2 \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F}), (8\sqrt{n})^{-1}\}$, so by eq. (1)

$$\begin{aligned}
&\log \mathcal{N}(\epsilon_k, \mathcal{L} \circ \mathcal{F}|_{\boldsymbol{z}}^r, \rho_{\boldsymbol{z},2}) \\
&\leq (\mathfrak{R}_{nq}(\Pi \circ \mathcal{F}))^2 \cdot \frac{4nq}{\xi_k^2} \cdot \log \frac{2e\beta nq}{\xi_k} \\
&\leq \left(2^{1+\theta}\lambda r^\theta \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F})\right)^2 \cdot \frac{4nq}{\epsilon_k^2} \cdot \log \left(e\beta(nq)^{3/2}\right) \\
&\leq \left(2^{1+\theta}\lambda r^\theta \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F})\right)^2 \cdot \frac{6nq}{\epsilon_k^2} \cdot \log\left(e\beta nq\right).
\end{aligned}$$

Note also that by construction $K \leq 4\log(e\beta nq)$.

By Theorem 4 and $\epsilon_{k-1} = 2 \cdot \epsilon_k$ we deduce that

$$\begin{aligned}
&\hat{\mathfrak{R}}_{\boldsymbol{z}}(\mathcal{L} \circ \mathcal{F}|_{\boldsymbol{z}}^r) \\
&\leq 2 \cdot \sum_{k=1}^K (\epsilon_k + \epsilon_{k-1}) \cdot \sqrt{\frac{\log \mathcal{N}(\epsilon_k, \mathcal{L} \circ \mathcal{F}|_{\boldsymbol{z}}^r, \rho_{\boldsymbol{z},2})}{n}} + \epsilon_K \\
&\leq 6 \sum_{k=1}^K \epsilon_k \cdot \sqrt{\frac{\log \mathcal{N}(\epsilon_k, \mathcal{L} \circ \mathcal{F}|_{\boldsymbol{z}}^r, \rho_{\boldsymbol{z},2})}{n}} + \epsilon_K \\
&\leq 6K \cdot \left(2^{1+\theta}\lambda r^\theta \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F})\right) \cdot \sqrt{6q \cdot \log\left(e\beta nq\right)} + \epsilon_K \\
&\leq 2^8\sqrt{q} \cdot \left(\lambda r^\theta \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F})\right) \cdot \log^{3/2}\left(e\beta nq\right) + \epsilon_K \\
&\leq \lambda r^\theta \left(2^9\sqrt{q} \cdot \log^{3/2}\left(e\beta nq\right) \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F}) + n^{-1/2}\right).
\end{aligned}$$

This completes the proof of the proposition. □

4.2. Proof of Theorem 1

To complete the proof of Theorem 1 we combine Proposition 1 with some results due to Bousquet (2002).

Theorem 5. *Suppose we have a measurable space $\mathcal{Z}$ and a function class $\mathcal{G} \subseteq \mathcal{M}(\mathcal{Z}, [0, b])$. For each $\boldsymbol{z} \in \mathcal{Z}^n$ and $g \in \mathcal{G}$ we let $\hat{\mathbb{E}}_{\boldsymbol{z}}(g) = n^{-1} \cdot \sum_{i \in [n]} g(z_i)$. Suppose we have a function $\phi_n : [0, \infty) \to [0, \infty)$ which is non-negative, non-decreasing, not identically zero, and $\phi_n(r)/\sqrt{r}$ is non-increasing. Suppose further that for all $\boldsymbol{z} \in \mathcal{Z}^n$ and $r > 0$,*

$$\hat{\mathfrak{R}}_{\boldsymbol{z}}(\{g \in \mathcal{G} : \hat{\mathbb{E}}_{\boldsymbol{z}}(g) \leq r\}) \leq \phi_n(r).$$

Let $\hat{r}_n$ be the largest solution of the equation $\phi_n(r) = r$. Suppose that Z is a random variable with distribution P, where P is a distribution on $\mathcal{Z}$ and let $\mathcal{D} = \{Z_i\}_{i \in [n]} \in \mathcal{Z}^n$ be an i.i.d. sample, where each $Z_i \sim P$ is an independent copy of Z. For any $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$, for all $g \in \mathcal{G}$:

$$\mathbb{E}(g) \leq \hat{\mathbb{E}}_{\mathcal{D}}(g) + 90(\hat{r}_n + r_0) + 4\sqrt{\hat{\mathbb{E}}_{\mathcal{D}}(g)(\hat{r}_n + r_0)}.$$

where $r_0 = b\,(\log(1/\delta) + 6\log\log n)/n$.

Proof. The following result is given in the penultimate line of the proof of (Theorem 6.1, Bousquet (2002)):

$$\mathbb{E}(g) \leq \hat{\mathbb{E}}_{\mathcal{D}}(g) + 45\hat{r}_n + \sqrt{8\hat{r}_n\mathbb{E}(g)} + \sqrt{4r_0 \cdot \mathbb{E}(g)} + 20r_0,$$

with probability at least $1 - \delta$, for all $g \in \mathcal{G}$. So,

$$\mathbb{E}(g) \leq \hat{\mathbb{E}}_{\mathcal{D}}(g) + 45\hat{r}_n + 20r_0 + 4\sqrt{(\hat{r}_n + r_0) \cdot \mathbb{E}(g)}.$$

We also need the following inequality (Lemma 5.11, Bousquet (2002)): Suppose that $t, B, C > 0$ satisfy $t \leq B\sqrt{t} + C$. Then $t \leq B^2 + C + B\sqrt{C}$. Applying this with $B = 4\sqrt{(\hat{r}_n + r_0)}$ and $C = \hat{\mathbb{E}}_{\mathcal{D}}(g) + 45\hat{r}_n + 20r_0$ we have

$$\begin{aligned}
\mathbb{E}(g) &\leq 16(\hat{r}_n + r_0) + (\hat{\mathbb{E}}_{\mathcal{D}}(g) + 45\hat{r}_n + 20r_0) \\
&\quad + 4\sqrt{(\hat{r}_n + r_0)(\hat{\mathbb{E}}_{\mathcal{D}}(g) + 45\hat{r}_n + 20r_0)} \\
&\leq \hat{\mathbb{E}}_{\mathcal{D}}(g) + 90(\hat{r}_n + r_0) + 4\sqrt{\hat{\mathbb{E}}_{\mathcal{D}}(g)(\hat{r}_n + r_0)},
\end{aligned}$$

which completes the proof. □

Theorem 5 is a uniform upper bound in terms of the empirical risk. We can deduce a performance bound on the empirical risk minimiser by combining with Bernstein's inequality – see Theorem 2.10 from (Boucheron et al., 2013)

Theorem 6 (Bernstein (1924))**.** *Let $W_i, \cdots, W_i \in [0, b]$ be bounded independent random variables with mean $\mu = \mathbb{E}[W_i]$. Then with probability at least $1 - \delta$ we have*

$$\begin{aligned}
\frac{1}{n}\sum_{i \in [n]} W_i &\leq \mu + \sqrt{\frac{2\mu b \log(1/\delta)}{n}} + \frac{b\log(1/\delta)}{n} \\
&\leq 2\mu + \frac{3b\log(1/\delta)}{2n}.
\end{aligned}$$

Corollary 1. *Suppose that the assumptions of Theorem 5 hold and choose $g^* \in \operatorname{argmin}_{g \in \mathcal{G}}\{\mathbb{E}(g)\}$. Given $\boldsymbol{z} \in \mathcal{Z}^n$ we choose $\hat{g}_{\boldsymbol{z}} \in \operatorname{argmin}_{g \in \mathcal{G}}\{\hat{\mathbb{E}}_{\boldsymbol{z}}(g)\}$. For any $\delta \in (0, 1)$, the following holds with probability at least $1 - 2\delta$*

$$\mathbb{E}(\hat{g}_{\mathcal{D}}) \leq \mathbb{E}(g^*) + 9\sqrt{\mathbb{E}(g^*) \cdot (\hat{r}_n + r_0)} + 100\,(\hat{r}_n + r_0).$$

We can now complete the proof of Theorem 1.

Proof of Theorem 1. First let $\mathcal{G} = \mathcal{L} \circ \mathcal{F} = \{(x, y) \mapsto \mathcal{L}(f(x), y) : f \in \mathcal{F}\}$. Note that for $g = \mathcal{L} \circ f$ with $f \in \mathcal{F}$ and $Z = (X, Y) \sim P$ we have $\mathbb{E}_Z(g) = \mathcal{E}_{\mathcal{L}}(f, P)$ and given $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ we have $\hat{\mathbb{E}}_{\boldsymbol{z}}(g) = \hat{\mathcal{E}}_{\mathcal{L}}(f)$. Note also that under this correspondence $\mathcal{L} \circ \mathcal{F}|^r_{\boldsymbol{z}} = \{g \in \mathcal{G} : \hat{\mathbb{E}}_{\boldsymbol{z}}(g) \leq r\}$.

Now define $\phi_n : [0, \infty) \to [0, \infty)$ by

$$\phi_n(r) = \lambda r^{\theta} \left(2^9 \sqrt{q} \cdot \log^{3/2}(e\beta nq) \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F}) + n^{-1/2}\right).$$

By Proposition 1, for each $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$,

$$\hat{\mathfrak{R}}_{\boldsymbol{z}}\left(\{g \in \mathcal{G} : \hat{\mathbb{E}}_{\boldsymbol{z}}(g) \leq r\}\right) = \hat{\mathfrak{R}}_{\boldsymbol{z}}\left(\mathcal{L} \circ \mathcal{F}|^r_{\boldsymbol{z}}\right) \leq \phi_n(r).$$

Observe that ϕ_n is non-negative, non-decreasing and $\phi_n(r)/\sqrt{r}$ is non-increasing, since $\theta \in [0, 1/2]$. So it remains to solve the fixed point equation $\phi_n(r) = r$ and find $\hat{r}_n := \left(\lambda\left(2^9\sqrt{q} \cdot \log^{3/2}(e\beta nq) \cdot \mathfrak{R}_{nq}(\Pi \circ \mathcal{F}) + n^{-1/2}\right)\right)^{\frac{1}{1-\theta}}$ Hence, the two bounds in Theorem 1 follow from Theorem 5 and Corollary 1, respectively. □

5. An application to ensembles

In this section we highlight applications of our general multi-output learning framework. Specifically, here we consider ensembles of decision trees (Schapire & Freund, 2013), as they represent an effective and widely used tool in practice (Chen & Guestrin, 2016).

Throughout this section we shall assume that $\mathcal{X} = \mathbb{R}^d$. We consider sets of decision tree functions $\mathcal{H}^1_{p,\tau} \subseteq \mathcal{M}(\mathcal{X}, [-1, 1]^q)$ constructed as follows. Let $\mathcal{T}_{p,d}$ be the set of decision trees $t : \mathbb{R}^d \to [p]$ with p leaves, where each internal node performs a binary split along a single feature. Let $\mathcal{H}_{p,\tau} \subseteq \mathcal{M}(\mathcal{X}, \mathbb{R}^q)$ be the set of all functions of the form $h(x) = (w_{t(x),j})_{j \in [q]}$, where $t \in \mathcal{T}_{p,d}$ is a decision tree and $\boldsymbol{w} = (w_{l,j})_{(l,j) \in [p] \times [q]} \in \mathbb{R}^{pq}$ satisfies the ℓ_1 constraint $\|w_{l\cdot}\|_1 = \sum_{j \in [q]} |w_{lj}| \leq \tau$. Finally, let $\mathcal{H}^1_{p,\tau} := \mathcal{H}_{p,\tau} \cap \mathcal{M}(\mathcal{X}, [-1, 1]^q)$. We now give a bound for convex combinations of such decision trees.

Theorem 7. *Suppose we have $\beta, b \geq 1$, $\lambda > 0$, $\theta \in [0, 1/2]$ and a (λ, θ)-self-bounding Lipschitz loss function $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0, b]$. Given $\delta \in (0, 1)$, $n \in \mathbb{N}$ we define for each $\alpha = (\alpha_t)_{t \in [T]}$, $\tau = (\tau_t)_{t \in [T]} \in (0, \infty)^T$,*

$$\mathfrak{C}_{n,\delta}(\alpha, \tau) := \left(\tfrac{\lambda}{\sqrt{n}}\left(\sqrt{p}\log^2(3npqd\beta) \cdot \textstyle\sum_{t \in [T]} \alpha_t \cdot \tau_t + 1\right)\right)^{\frac{1}{1-\theta}} + \tfrac{b}{n} \cdot (\log(1/\delta) + \log(\log n)).$$

There exists a numerical constant C_0 such that given an i.i.d. sample $\mathcal{D}$ the following holds with probability at least $1 - \delta$, for all ensembles $f = \sum_{t \in [T]} \alpha_t \cdot h_t$ where $\sum_{t \in [T]} \alpha_t \leq \beta$ and $h_t \in \mathcal{H}^1_{p,\tau_t}$,

$$\mathcal{E}_{\mathcal{L}}(f) \leq \hat{\mathcal{E}}_{\mathcal{L}}(f) + C_0 \cdot \left(\sqrt{\hat{\mathcal{E}}_{\mathcal{L}}(f) \cdot \mathfrak{C}_{n,\delta}(\alpha, \tau)} + \mathfrak{C}_{n,\delta}(\alpha, \tau)\right).$$

Before giving the proof, we highlight several interesting features of this result:

- First and foremost, Theorem 7 gives guarantees for ensembles of decision trees with respect to a wide variety of losses including the *multinomial logistic loss* for multi-class classification and the *one versus all loss* for multi-label classification, as well as implying margin based guarantees (see Section 2.2).
- Theorem 7 has a favourable dependency upon the number of examples whenever $\hat{\mathcal{E}}_{\mathcal{L}}(f)$ is sufficiently small, as is often the case for large ensembles of decision trees. For example, if we are using the multinomial logistic loss and $\hat{\mathcal{E}}_{\mathcal{L}}(f) \approx 0$, then Theorem 7 gives rise to a fast rate of $O(n^{-1})$.
- Theorem 7 has only logarithmic dependency upon the dimensionality of the output space q. This contrasts starkly with previous guarantees for multi-class learning with ensembles of decision trees (Kuznetsov et al., 2014; 2015) which are linear with respect to the number of classes q.

5.1. Proof of Theorem 7

The proof of Theorem 7 is a consequence of Theorem 1 combined with the following lemma.

Lemma 5. *For all $m \in \mathbb{N}$ and $\boldsymbol{z} \in (\mathcal{X} \times [q])^m$ we have,*

$$\hat{\mathfrak{R}}_{\boldsymbol{z}}(\Pi \circ \mathcal{H}_{p,\tau}) \leq 2\tau \cdot \sqrt{p \cdot \log(2 \cdot \max\{p \cdot d \cdot m, q\})/m}.$$

We begin by counting the number of possible partitions that can be made by a decision tree in $\mathcal{T}_{p,d}$ on a given sequence of points. Given a sequence $\boldsymbol{x} = (x_i)_{i \in [m]} \in \mathcal{X}^m$ we let $\mathcal{T}_{p,d}(\boldsymbol{x}) := \left\{(t(x_i))_{i \in [m]} : t \in \mathcal{T}_{p,d}\right\} \subseteq [p]^m$.

Lemma 6. *For all $m \in \mathbb{N}$ and $\boldsymbol{x} \in \mathcal{X}^m$, we have $|\mathcal{T}_{p,d}(\boldsymbol{x})| \leq (p-1)! \cdot (d \cdot (m+1))^{p-1}$.*

Proof. By induction, it suffices to show that $|\mathcal{T}_{p+1,d}(\boldsymbol{x})| \leq |\mathcal{T}_{p,d}(\boldsymbol{x})| \cdot p \cdot d \cdot (m+1)$. Now observe that each element of $\mathcal{T}_{p+1,d}(\boldsymbol{x})$ may be constructed by taking an element of $\mathcal{T}_{p,d}(\boldsymbol{x})$ and then making a choice of one of p existing leaf nodes to partition, one of d dimensions to split upon, and one of at most $m+1$ possible split points. □

We complete the proof of Lemma 5 as follows.

Proof of Lemma 5. For a given $\tau > 0$ we let $\Lambda_\tau := \{(a_j)_{j \in [q]} : \sum_{j \in [q]} |a_j| \leq \tau\}$. Let $\{e(j)\}_{j \in [q]} \subseteq \mathbb{R}^q$ be the canonical orthonormal basis. Let $\Lambda^{\text{ex}}_1 \subseteq \Lambda_1$ denote the subset of extreme points in Λ_1, so $\Lambda^{\text{ex}}_1 = \{u \cdot e(j) : u \in \{-1, +1\} \text{ and } j \in [q]\}$. Note that $|\Lambda^{\text{ex}}_1| = 2q$. Fix

$\boldsymbol{z} = (z_i)_{i\in[m]} \in (\mathcal{X}\times[q])^m$ with each $z_i = (x_i, j_i)$ and let $\boldsymbol{x} = (x_i)_{i\in[m]} \in \mathcal{X}^m$. Then $\hat{\mathfrak{R}}_{\boldsymbol{z}}(\Pi\circ\mathcal{H}_{p,\tau}) =$

$$
\begin{aligned}
&= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{h\in\mathcal{H}_{p,\tau}} \left\{ \frac{1}{m}\sum_{i\in[m]} \sigma_i \cdot (\Pi\circ h)(x_i, j_i) \right\} \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{h\in\mathcal{H}_{p,\tau}} \left\{ \frac{1}{m}\sum_{i\in[m]} \sigma_i \cdot \pi_{j_i}(h(x_i)) \right\} \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{t\in\mathcal{T}_{p,d}} \left\{ \sup_{w\in(\Lambda_\tau)^p} \left\{ \frac{1}{m}\sum_{i\in[m]} \sigma_i \cdot w_{t(x_i),j_i} \right\} \right\} \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{(l_i)_{i\in[m]}\in\mathcal{T}_{p,d}(\boldsymbol{x})} \left\{ \frac{1}{m} \cdot \sup_{w\in(\Lambda_\tau)^p} \left\{ \sum_{i\in[m]} \sigma_i \cdot w_{l_i,j_i} \right\} \right\}
\end{aligned}
$$

Observe that for $w \in (\Lambda_\tau)^p$ and $(l_i) \in [p]^m$,

$$
\begin{aligned}
&\sum_{i\in[m]} \sigma_i \cdot w_{l_i,j_i} \\
&= \sum_{r\in[p]}\sum_{s\in[q]} w_{r,s} \sum_{i:l_i=r\,\&\,j_i=s} \sigma_i \\
&\le \sum_{r\in[p]} \left\{ \left(\sum_{s\in[q]} |w_{r,s}| \right) \cdot \max_{s\in[q]} \left| \sum_{i:l_i=r\,\&\,j_i=s} \sigma_i \right| \right\} \\
&\le \tau \cdot \sum_{r\in[p]} \max_{s\in[q]} \left| \sum_{i:l_i=r\,\&\,j_i=s} \sigma_i \right| \\
&= \tau \cdot \sum_{r\in[p]} \sup_{(u_{r,s})_{s\in[q]}\in\Lambda_1^{\mathrm{ex}}} \left\{ \sum_{s\in[q]} u_{r,s} \left(\sum_{i:l_i=r\,\&\,j_i=s} \sigma_i \right) \right\} \\
&= \tau \cdot \sup_{(u_{r,s})\in(\Lambda_1^{\mathrm{ex}})^p} \left\{ \sum_{r\in[p]}\sum_{s\in[q]} u_{r,s} \left(\sum_{i:l_i=r\,\&\,j_i=s} \sigma_i \right) \right\} \\
&= \tau \cdot \sup_{(u_{r,s})\in(\Lambda_1^{\mathrm{ex}})^p} \left\{ \sum_{i\in[m]} \sigma_i \cdot u_{l_i,j_i} \right\}.
\end{aligned}
$$

Plugging this bound into the above yields $\hat{\mathfrak{R}}_{\boldsymbol{z}}(\Pi\circ\mathcal{H}_{p,\tau})$

$$
\begin{aligned}
&\le \mathbb{E}_{\boldsymbol{\sigma}} \sup_{(l_i)_{i\in[m]}\in\mathcal{T}_{p,d}(\boldsymbol{x})} \left\{ \frac{\tau}{m} \sup_{(u_{r,s})\in(\Lambda_1^{\mathrm{ex}})^p} \left\{ \sum_{i\in[m]} \sigma_i \cdot u_{l_i,j_i} \right\} \right\} \\
&\le \tau \cdot \sup_{(l_i)_{i\in[m]}\in\mathcal{T}_{p,d}(\boldsymbol{x}),\ (u_{r,s})\in(\Lambda_1^{\mathrm{ex}})^p} \left\{ \sqrt{\sum_{i\in[m]} u_{l_i,j_i}^2} \right\} \cdot \ldots \\
&\qquad \cdot \frac{\sqrt{2\log\left(|\mathcal{T}_{p,d}(\boldsymbol{x})| \cdot |\Lambda_1^{\mathrm{ex}}|^p\right)}}{m} \qquad (2) \\
&\le \tau \cdot \sqrt{\frac{2\left((p-1)\log(p\cdot d\cdot m) + p\cdot\log(2q)\right)}{m}} \\
&\le 2\tau \cdot \sqrt{\frac{p\cdot\log(2\cdot\max\{p\cdot d\cdot m, q\})}{m}},
\end{aligned}
$$

where (2) follows from Massart's lemma (eg. Theorem 3.3 from (Mohri et al., 2012)) and the penultimate inequality follows from Lemma 6. □

We can now deduce Theorem 7 from Theorem 1 with the help of a re-weighting argument along with the convexity property of Rademacher complexities.

Proof of Theorem 7. Take $\zeta > 0$ and let

$$
\mathcal{F} := \left\{ f = \sum_{t\in[T]} \alpha_t \cdot h_t : \quad h_t \in \mathcal{H}^1_{p,\tau_t},\quad \alpha_t \ge 0, \right.
$$
$$
\left. \sum_{t\in[T]} \alpha_t\cdot\tau_t \le \zeta \text{ and } \sum_{t\in[T]} \alpha_t \le \beta \right\}.
$$

Observe that $\mathcal{F} \subseteq \mathrm{conv}\,(\mathcal{H}_{p,\zeta})$. Indeed, given $f = \sum_{t\in[T]} \alpha_t \cdot h_t$ with $h_t \in \mathcal{H}^1_{p,\tau_t}$ and $\sum_{t\in[T]} \alpha_t\cdot\tau_t \le \zeta$, we can rewrite

$$
f = \sum_{t\in[T]} \left(\frac{\alpha_t\cdot\tau_t}{\zeta} \right) \cdot (\zeta\cdot {\tau_t}^{-1} \cdot h_t),
$$

with $\sum_{t\in[T]} (\alpha_t\cdot\tau_t\cdot\zeta^{-1}) \le 1$ and for each $t\in[T]$, we have $\zeta\cdot{\tau_t}^{-1}\cdot h_t \in \mathcal{H}_{p,\zeta}$. Thus, $\Pi\circ\mathcal{F} \subseteq \Pi\circ\mathrm{conv}\,(\mathcal{H}_{p,\zeta}) = \mathrm{conv}\,(\Pi\circ\mathcal{H}_{p,\zeta})$. Hence, by the convexity of Rademacher complexities (Boucheron et al., 2005, Theorem 3.3, eq. (5)) and $\boldsymbol{z} \in (\mathcal{X}\times[q])^{nq}$ combined with Lemma 5 we have,

$$
\begin{aligned}
\hat{\mathfrak{R}}_{\boldsymbol{z}}(\Pi\circ\mathcal{F}) \le \hat{\mathfrak{R}}_{\boldsymbol{z}}(\mathrm{conv}\,(\Pi\circ\mathcal{H}_{p,\zeta})) &\le \hat{\mathfrak{R}}_{\boldsymbol{z}}(\Pi\circ\mathcal{H}_{p,\zeta}) \\
&\le 2\zeta\cdot\sqrt{\frac{p\cdot\log(2\cdot\max\{pd\cdot(nq), q\})}{nq}} \\
&= 2\zeta\cdot\sqrt{\frac{p\cdot\log(2pdnq)}{nq}}.
\end{aligned}
$$

Taking a supremum over all $\boldsymbol{z} \in (\mathcal{X}\times[q])^{nq}$ we have $\mathfrak{R}_{nq}(\Pi\circ\mathcal{F}) \le 2\zeta\cdot\sqrt{(p\cdot\log(2pdnq))(nq)^{-1}}$. Note also that $\mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, [-\beta,\beta]^q)$, since each $f\in\mathcal{F}$ is of the form $f = \sum_{t\in[T]} \alpha_t\cdot h_t$ with $h_t \in \mathcal{H}^1_{p,\tau_t} \subseteq \mathcal{M}(\mathcal{X}, [-1,+1]^q)$ and $\sum_{t\in[T]} \alpha_t \le \beta$. Thus, plugging the bound on $\mathfrak{R}_{nq}(\Pi\circ\mathcal{F})$ into Theorem 1 yields the bound in Theorem 7. □

6. Conclusions

We presented a theoretical analysis of multi-output learning, based on a self-bounding Lipschitz condition. Under this condition, we obtained favourable dependence on both the sample size and the output dimension. The main analytic tool is a new contraction inequality for the local Rademacher complexity of vector valued function classes with a self-bounding Lipschitz loss, which may be of independent interest. Theorem 1 can be applied to any multi-output prediction problem where one can obtain an upper bound on the Rademacher complexity $\mathfrak{R}_{nq}(\Pi\circ\mathcal{F})$. We demonstrate this by applying our approach to ensembles of decision trees, yielding state of the art results.

Acknowledgement

This work is funded by EPSRC (grant EP/P004245/1) and partially by the Turing Institute (grant EP/N510129/1).

References

Agrawal, R., Gupta, A., Prabhu, Y., and Varma, M. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 13–24, 2013.

Babbar, R. and Schölkopf, B. Dismec: Distributed sparse machines for extreme multi-label classification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 721–729, 2017.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

Bernstein, S. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

Bhatia, K., Jain, H., Kar, P., Varma, M., and Jain, P. Sparse local embeddings for extreme multi-label classification. In *Advances in neural information processing systems*, pp. 730–738, 2015.

Borchani, H., Varando, G., Bielza, C., and Larrañaga, P. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5): 216–233, 2015.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: A survey of some recent advances, 2005.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Bousquet, O. Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms. *PhD Thesis*, 2002.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Chzhen, E. Classification of sparse binary vectors. *arXiv preprint arXiv:1903.11867*, 2019.

Chzhen, E., Denis, C., Hebiri, M., and Salmon, J. On the benefits of output sparsity for multi-label classification. *arXiv preprint arXiv:1703.04697*, 2017.

Cortes, C., Kuznetsov, V., Mohri, M., and Yang, S. Structured prediction theory based on factor graph complexity. In *International Conference on Machine Learning*, pp. 2522–2530, 2016.

Crammer, K. and Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

Dudley, R. M. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Geng, X. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

Guermeur, Y. Lp-norm sauer–shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences*, 89:450–473, 2017.

Jain, H., Balasubramanian, V., Chunduri, B., and Varma, M. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 528–536, 2019.

Koltchinskii, V. et al. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Kuznetsov, V., Mohri, M., and Syed, U. Multi-class deep boosting. In *Advances in Neural Information Processing Systems*, pp. 2501–2509, 2014.

Kuznetsov, V., Mohri, M., and Syed, U. Rademacher complexity margin bounds for learning with a large number of classes. In *ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels*, 2015.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Lei, Y., Dogan, U., Binder, A., and Kloft, M. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In *Advances in Neural Information Processing Systems*, pp. 2035–2043, 2015.

Lei, Y., Ding, L., and Bi, Y. Local rademacher complexity bounds based on covering numbers. *Neurocomputing*, 218:320–330, 2016.

Lei, Y., Dogan, Ü., Zhou, D.-X., and Kloft, M. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5): 2995–3021, 2019.

Li, J., Liu, Y., Yin, R., Zhang, H., Ding, L., and Wang, W. Multi-class learning: From theory to algorithm. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1586–1595. Curran Associates, Inc., 2018.

Li, J., Liu, Y., Yin, R., and Wang, W. Multi-class learning using unlabeled samples: theory and algorithm. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 2880–2886. AAAI Press, 2019.

Liu, Y., Li, J., Ding, L., Liu, X., and Wang, W. Learning vector-valued functions with local rademacher complexity and unlabeled data, 2019.

Mammen, E. and Tsybakov, A. B. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 12 1999. doi: 10.1214/aos/1017939240.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17. Springer, 2016.

Menon, A. K., Rawat, A. S., Reddi, S., and Kumar, S. Multilabel reductions: what is my loss optimising? In *Advances in Neural Information Processing Systems*, pp. 10599–10610, 2019.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2012.

Musayeva, K., Lauer, F., and Guermeur, Y. Rademacher complexity and generalization performance of multi-category margin classifiers. *Neurocomputing*, pp. 6–15, 11 2019.

Reddi, S. J., Kale, S., Yu, F., Holtmann-Rice, D., Chen, J., and Kumar, S. Stochastic negative mining for learning with large output spaces. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1940–1949, 2019.

Reeve, H. W. and Kabán, A. Optimistic bounds for multi-output prediction, 2020.

Schapire, R. E. and Freund, Y. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pp. 2199–2207, 2010.

Talagrand, M. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

Tsoumakas, G. and Katakis, I. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

Xu, C., Liu, T., Tao, D., and Xu, C. Local rademacher complexity for multi-label learning. *IEEE Transactions on Image Processing*, 25(3):1495–1507, March 2016.

Xu, D., Shi, Y., Tsang, I. W., Ong, Y.-S., Gong, C., and Shen, X. Survey on multi-output learning. *IEEE transactions on neural networks and learning systems*, 2019.

Zhang, M.-L. and Zhou, Z.-H. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

Supplementary material for 'Optimistic bounds for multi-output prediction'

A. The self-bounding Lipschitz condition

A.1. Proof of Lemma 1

The proof of Lemma 1 starts with the following lemma.

Lemma 7. *Suppose that $\varphi : \mathbb{R} \to [0, \infty)$ is a non-negative differentiable function satisfying:*

1. *The derivative $\varphi'(t)$ is non-negative on $[0, \infty)$;*
2. *$\forall t_0, t_1 > 0$, $|\varphi'(t_1) - \varphi'(t_0)| \leq \left(\frac{\lambda}{2}\right)^{\frac{1}{1-\theta}} \cdot |t_1 - t_0|^{\frac{\theta}{1-\theta}}$.*

Then $\forall t > 0$, $\varphi'(t) \leq \lambda \cdot \varphi(t)^\theta$. Moreover, for all $t > 0$, $\varphi(t) - \varphi(0) \leq \lambda \cdot \varphi(t)^\theta \cdot t$.

Proof. Fix $t > 0$ and take $\Delta = 2\lambda^{-\frac{1}{\theta}} \cdot \varphi'(t)^{\frac{1-\theta}{\theta}}$, which is positive by the first condition. By the non-negativity of φ and the mean value theorem there exists some $s \in (t - \Delta, t)$

$$\begin{aligned}
0 \leq \varphi(t - \Delta) &\leq \varphi(t) - \varphi'(s) \cdot \Delta \\
&\leq \varphi(t) - \varphi'(t) \cdot \Delta + |\varphi'(s) - \varphi'(t)| \cdot \Delta \\
&\leq \varphi(t) - \varphi'(t) \cdot \Delta + \left((\lambda/2)^{\frac{1}{1-\theta}} \cdot \Delta^{\frac{\theta}{1-\theta}}\right) \cdot \Delta \\
&\leq \varphi(t) - \varphi'(t) \cdot \Delta + (\lambda \cdot \Delta/2)^{\frac{1}{1-\theta}} \\
&\leq \varphi(t) - 2(\varphi'(t)/\lambda)^{\frac{1}{\theta}} + (\varphi'(t)/\lambda)^{\frac{1}{\theta}} \\
&= \varphi(t) - (\varphi'(t)/\lambda)^{\frac{1}{\theta}},
\end{aligned}$$

where the fourth inequality follows from the second condition. Rearranging completes the proof of the first part of the lemma.

To prove the second part of the lemma we apply the mean value theorem combined with the first part of the lemma to obtain for some $s \in (0, t)$,

$$\varphi(t) - \varphi(0) = \varphi'(s) \cdot t \leq \left(\lambda \cdot \varphi(s)^\theta\right) \cdot t \leq \lambda \cdot \varphi(t)^\theta \cdot t,$$

where we used the non-negativity of φ' on $[0, \infty)$ to ensure that $\varphi(s) \leq \varphi(t)$. This completes the proof of the lemma. □

Proof of Lemma 1. Take $u, v \in \mathcal{V}$ and $y \in \mathcal{Y}$. Without loss of generality we assume that $\mathcal{L}(u, y) \leq \mathcal{L}(v, y)$ and let $\varphi_{u,y}$ be a function satisfying the conditions specified in the statement of the lemma. By combining the first two conditions with Lemma 7 we see that $\varphi_{u,y}(t) - \varphi_{u,y}(0) \leq \lambda \cdot \varphi_{u,y}(t)^\theta \cdot t$. Hence, by dividing through by $\varphi_{u,y}(t)^\theta$ and applying $\mathcal{L}(v, y) \leq \varphi_{u,y}(t)$ twice we have,

$$\begin{aligned}
\mathcal{L}(v, y)^{1-\theta} - \lambda \cdot t &\leq \varphi_{u,y}(t)^{1-\theta} - \lambda \cdot t \\
&\leq \varphi_{u,y}(0) \cdot \varphi_{u,y}(t)^{-\theta} \\
&\leq \varphi_{u,y}(0) \cdot \mathcal{L}(v, y)^{-\theta} \\
&= \mathcal{L}(u, y) \cdot \mathcal{L}(v, y)^{-\theta}.
\end{aligned}$$

Multiplying by $\mathcal{L}(v, y)^\theta$ and rearranging we have $\mathcal{L}(v, y) - \mathcal{L}(u, y) \leq \lambda \cdot \mathcal{L}(v, y)^\theta$. Since $\mathcal{L}(u, y) \leq \mathcal{L}(v, y)$ this completes the proof of the lemma. □

A.2. Proof of Lemma 2

Proof of Lemma 2. Take $u, v \in \mathcal{V}$ and $y \in \mathcal{Y}$. Without loss of generality we assume that $\tilde{\mathcal{L}}(u, y) \leq \tilde{\mathcal{L}}(v, y)$, so it suffices to show that

$$\tilde{\mathcal{L}}(v, y) - \tilde{\mathcal{L}}(u, y) \leq \lambda \cdot \tilde{\mathcal{L}}(v, y)^\theta \cdot \|u - v\|_\infty. \tag{3}$$

If $\mathcal{L}(u,y) \geq b$ then $\tilde{\mathcal{L}}(v,y) = \tilde{\mathcal{L}}(u,y) = b$, so (3) clearly holds. Thus, we can assume $\mathcal{L}(u,y) < b$, so $\tilde{\mathcal{L}}(u,y) = \mathcal{L}(u,y)$. By the (λ,θ) self-bounding Lipschitz condition for $\mathcal{L}$ we have

$$\begin{aligned}\mathcal{L}(v,y) - \tilde{\mathcal{L}}(u,y) &= \mathcal{L}(v,y) - \mathcal{L}(u,y) \\ &\leq \lambda \cdot \mathcal{L}(v,y)^{\theta} \cdot \|u-v\|_{\infty}.\end{aligned}$$

Equivalently, we have

$$\mathcal{L}(v,y)^{1-\theta} - \lambda \cdot \|u-v\|_{\infty} \leq \tilde{\mathcal{L}}(u,y) \cdot \mathcal{L}(v,y)^{-\theta}.$$

Since $\tilde{\mathcal{L}}(v,y) \leq \mathcal{L}(v,y)$, we deduce

$$\tilde{\mathcal{L}}(v,y)^{1-\theta} - \lambda \cdot \|u-v\|_{\infty} \leq \tilde{\mathcal{L}}(u,y) \cdot \tilde{\mathcal{L}}(v,y)^{-\theta}.$$

Rearranging gives (3) and completes the proof of the lemma. □

A.3. Proof of Proposition 2

The following result shows an example application of Lemma 1. We may verify the self-bounding Lipschitz condition for other loss functions in a similar manner.

Proposition 2. *Take $\mathcal{Y} = [q]$ and define the multinomial logistic loss $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to [0,\infty)$ is defined by*

$$\mathcal{L}(u,y) = \log\Big(\sum_{j\in[q]} \exp(u_j - u_y)\Big),$$

*where $u = (u_j)_{j\in[q]}$ and $y \in [q]$. It follows that $\mathcal{L}$ is (λ,θ)-*self-bounding Lipschitz *with $\lambda = 1$ and $\theta = 1/2$.*

The proof of Proposition 2 requires the following elementary lemma.

Lemma 8. *Given any $A > 0$ the function $\varphi : \mathbb{R} \to (0,\infty)$ defined by $\varphi(t) = \log(1 + A \cdot \exp(2t))$ is differentiable $\varphi'(t_0) > 0$ and and $|\varphi'(t_0) - \varphi'(t_1)| \leq |t_1 - t_0|$ for all $t_0, t_1 \in \mathbb{R}$.*

Proof. We begin by computing the first three derivatives,

$$\begin{aligned}\varphi'(t) &= \frac{2A \cdot \exp(2t)}{1 + A \cdot \exp(2t)} \\ \varphi''(t) &= \frac{4A \cdot \exp(2t)}{(1 + A \cdot \exp(2t))^2} \\ \varphi'''(t) &= \frac{8A \cdot \exp(2t)}{(1 + A \cdot \exp(2t))^3} \cdot (1 - A \cdot \exp(2t)).\end{aligned}$$

Clearly we have $\varphi'(t), \varphi''(t) > 0$ for all $t \in \mathbb{R}$. Moreover, by inspecting the third derivative we see that φ'' has a unique maximum where $A \cdot \exp(2t) = 1$. This implies that φ is twice differentiable with $|\varphi''(t)| \leq 1/4$ for all $t \in \mathbb{R}$. By the mean value theorem this yields $|\varphi'(t_0) - \varphi'(t_1)| \leq |t_1 - t_0|$ for all $t_0, t_1 \in \mathbb{R}$. □

Proof of Proposition 2. To complete the proof we $A_{u,y} := \sum_{j\in[q]\setminus\{y\}} \exp(u_j - u_y)$ and define $\varphi_{u,y}(t) := \log(1 + A_{u,y} \cdot \exp(2t))$. We can apply Lemma 8 to confirm that $\varphi_{u,y}$ satisfies the conditions of Lemma 1. Hence, the conclusion of Proposition 2 follows from Lemma 1. □

B. Proof of Theorem 2

For completeness here we give a proof of Theorem 2, which may be viewed as a mild generalisation of Theorem 6 from (Lei et al., 2019). We use the following well known result.

Theorem 8 ((Bartlett & Mendelson, 2002)). *Suppose we have a measurable space $\mathcal{Z}$ and a function class $\mathcal{G} \subseteq \mathcal{M}(\mathcal{Z}, [0, b])$. For each $\boldsymbol{z} \in \mathcal{Z}^n$ and $g \in \mathcal{G}$ we let $\hat{\mathbb{E}}_{\boldsymbol{z}}(g) = n^{-1} \cdot \sum_{i\in[n]} g(z_i)$. Suppose that Z is a random variable with distribution P is a distribution on $\mathcal{Z}$ and let $\mathcal{D} = \{Z_i\}_{i\in[n]} \in \mathcal{Z}^n$ be an i.i.d. which each $Z_i \sim P$, an independent copy of Z. For any $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$, for all $g \in \mathcal{G}$,*

$$\left|\mathbb{E}_Z(g) - \hat{\mathbb{E}}_{\boldsymbol{z}}(g)\right| \leq 2\mathbb{E}_{\mathcal{D}}\left[\hat{\mathfrak{R}}_{\mathcal{D}}\left(\mathcal{G}\right)\right] + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof of Theorem 2. With the correspondence introduced in the proof of Theorem 1, Theorem 8 implies that with probability at least $1 - \delta$ over a sample $\mathcal{D} = \{(X_i, Y_i)\}_{i\in[n]}$ with $(X_i, Y_i) \sim P$ the following holds for all $f \in \mathcal{F}$,

$$\left|\mathcal{E}_{\mathcal{L}}(f, P) - \hat{\mathcal{E}}_{\mathcal{L}}(f, \mathcal{D})\right| \leq 2\mathbb{E}_{\mathcal{D}}\left[\hat{\mathfrak{R}}_{\mathcal{D}}\left(\mathcal{L} \circ \mathcal{F}\right)\right] + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Hence, the result follows from Proposition 1 by taking $r = b$ and $\theta = 0$. □